Final

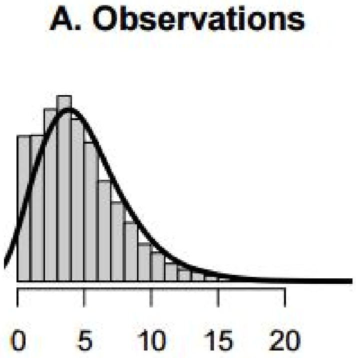
Christophe Hunt

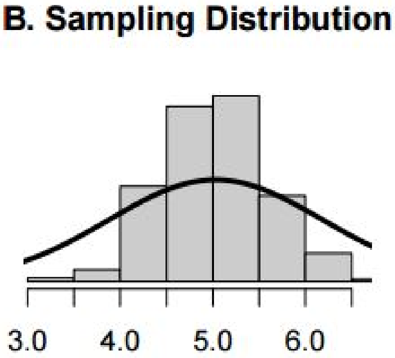
December 14, 2015

IS 606 Fall 2015 - Final Exam

Part I

Figure A below represents the distribution of an observed variable. Figure B below represents the distribution of the mean from 500 random samples of size 30 from A. The mean of A is 5.05 and the mean of B is 5.04. The standard deviations of A and B are 3.22 and 0.58, respectively


A. Observations


B. Sampling Distribution

a. Describe the two distributions.

The `Figure A` distribution is unimodal and right skewed. As previously stated the mean is 5.05 and this mean is to the right of the mode.

The `Figure B` distribution is unimodal and symetrical. As previously stated the mean is 5.04 and the distribution is symetrically distributed around the mean.

b. Explain why the means of these two distributions are similar but the standard deviations are not.

The means are similar because `Figure B` is a sampling distribution of `Figure A`. The expectation is that `Figure B` would have approximately the same mean as `Figure A` because `Figure B` is drawing from the population of Figure A. However, the standard deviations are dissimilar because of the distribution of observed values. In `Figure A` we have observations as high as 15 and as low as 0 which means there is a large variances in observable values. Whereas `Figure B` has a much more narrow distribution of values from 3 to approximately 7.

c. What is the statistical principal that describes this phenomenon?

The statistical principle at play here is the Central Limit Theorem. We expect the sampling distribution to be normally distributed around a mean that would be very closely match the mean of the original population.

Part II

Consider the four datasets, each with two columns (x and y), provided below.

```
options(digits = 2)
data1 <- data.frame(x = c(10, 8, 13, 9, 11, 14, 6, 4, 12, 7, 5), y = c(8.04,
    6.95, 7.58, 8.81, 8.33, 9.96, 7.24, 4.26, 10.84, 4.82, 5.68))
data2 <- data.frame(x = c(10, 8, 13, 9, 11, 14, 6, 4, 12, 7, 5), y = c(9.14,
    8.14, 8.74, 8.77, 9.26, 8.1, 6.13, 3.1, 9.13, 7.26, 4.74))
data3 <- data.frame(x = c(10, 8, 13, 9, 11, 14, 6, 4, 12, 7, 5), y = c(7.46,
    6.77, 12.74, 7.11, 7.81, 8.84, 6.08, 5.39, 8.15, 6.42, 5.73))
data4 <- data.frame(x = c(8, 8, 8, 8, 8, 8, 8, 19, 8, 8, 8), y = c(6.58, 5.76,
    7.71, 8.84, 8.47, 7.04, 5.25, 12.5, 5.56, 7.91, 6.89))
```

For each column, calculate (to two decimal places):

a. The mean (for x and y separately).

```
mean_data1 <- c(mean(data1$x), mean(data1$y))
mean_data2 <- c(mean(data2$x), mean(data2$y))
mean_data3 <- c(mean(data3$x), mean(data3$y))
mean_data4 <- c(mean(data4$x), mean(data4$y))
mean_table <- t(data.frame(mean_data1, mean_data2, mean_data3, mean_data4))
colnames(mean_table) <- c("x", "y")
kable(mean_table, caption = "Mean Table")
```

Table 1: Mean Table

	x	y
mean_data1	9	7.5
mean_data2	9	7.5
mean_data3	9	7.5
mean_data4	9	7.5

b. The median (for x and y separately).

```r
median_data1 <- c(median(data1$x), median(data1$y))
median_data2 <- c(median(data2$x), median(data2$y))
median_data3 <- c(median(data3$x), median(data3$y))
median_data4 <- c(median(data4$x), median(data4$y))
median_table <- t(data.frame(median_data1, median_data2, median_data3, median_data4))
colnames(median_table) <- c("x", "y")
kable(median_table, caption = "Median Table")
```

Table 2: Median Table

	x	y
median_data1	9	7.6
median_data2	9	8.1
median_data3	9	7.1
median_data4	8	7.0

c. The standard deviation (for x and y separately).

```r
sd_data1 <- c(sd(data1$x), sd(data1$y))
sd_data2 <- c(sd(data2$x), sd(data2$y))
sd_data3 <- c(sd(data3$x), sd(data3$y))
sd_data4 <- c(sd(data4$x), sd(data4$y))
sd_table <- t(data.frame(sd_data1, sd_data2, sd_data3, sd_data4))
colnames(sd_table) <- c("x", "y")
kable(sd_table, caption = "Standard Deviation Table")
```

Table 3: Standard Deviation Table

	x	y
sd_data1	3.3	2
sd_data2	3.3	2
sd_data3	3.3	2
sd_data4	3.3	2

For each x and y pair, calculate (also to two decimal places):

d. The correlation.

```
cor_data1 <- cor(data1$x, data1$y)
cor_data2 <- cor(data2$x, data2$y)
cor_data3 <- cor(data3$x, data3$y)
cor_data4 <- cor(data4$x, data4$y)
kable(t(data.frame(cor_data1, cor_data2, cor_data3, cor_data4)),
      caption = "Correlation Table")
```

Table 4: Correlation Table

cor_data1	0.82
cor_data2	0.82
cor_data3	0.82
cor_data4	0.82

e. Linear regression equation.

```
lmdata1 <- lm(x~y, data = data1)
lmdata2 <- lm(x~y, data = data2)
lmdata3 <- lm(x~y, data = data3)
lmdata4 <- lm(x~y, data = data4)

stargazer(lmdata1, lmdata2, lmdata3, lmdata4,
          title="Linear Regression Results",
          header = FALSE, dep.var.caption = "")
```

Table 5: Linear Regression Results

	x			
	(1)	(2)	(3)	(4)
y	1.300***	1.300***	1.300***	1.300***
	(0.310)	(0.310)	(0.320)	(0.310)
Constant	−1.000	−0.990	−1.000	−1.000
	(2.400)	(2.400)	(2.400)	(2.400)
Observations	11	11	11	11
R^2	0.670	0.670	0.670	0.670
Adjusted R^2	0.630	0.630	0.630	0.630
Residual Std. Error (df = 9)	2.000	2.000	2.000	2.000
F Statistic (df = 1; 9)	18.000***	18.000***	18.000***	18.000***
Note:		*p<0.1; **p<0.05; ***p<0.01		

f. R-Squared

```r
summarylmdata1 <- summary.lm(lmdata1)
summarylmdata2 <- summary.lm(lmdata2)
summarylmdata3 <- summary.lm(lmdata3)
summarylmdata4 <- summary.lm(lmdata4)

kable(t(data.frame(summarylmdata1$r.squared,
                   summarylmdata2$r.squared,
                   summarylmdata3$r.squared,
                   summarylmdata4$r.squared)),
      caption = "$R^2$ Results from Linear Regression")
```

Table 6: $R^2$ Results from Linear Regression

summarylmdata1.r.squared	0.67
summarylmdata2.r.squared	0.67
summarylmdata3.r.squared	0.67
summarylmdata4.r.squared	0.67

For each pair, is it appropriate to estimate a linear regression model? Why or why not? Be specific as to why for each pair! Explain why it is important to include appropriate visualizations when analyzing data. Include any visualization(s) you create.

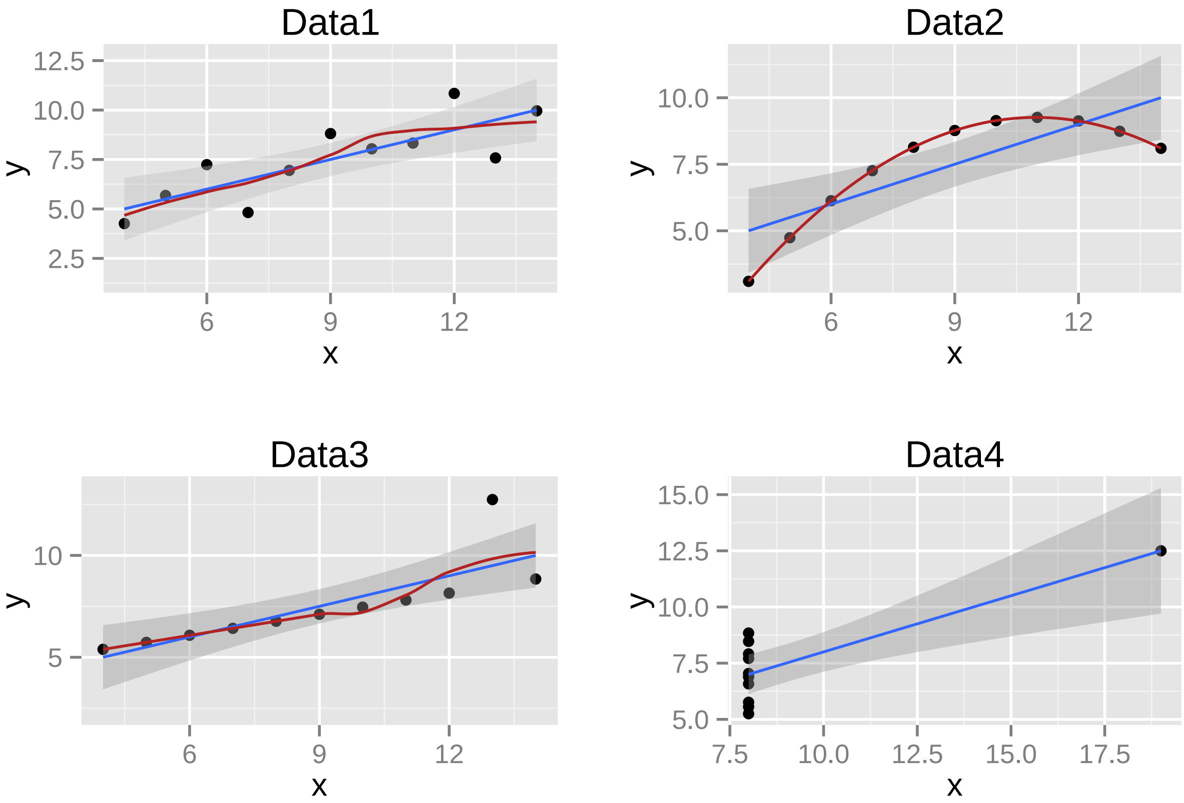
```r
plot_data1 <- ggplot(aes(x, y), data = data1) + geom_point() + geom_smooth(method = "lm",
    formula = y ~ x, fill = "grey") + stat_smooth(colour = "firebrick", alpha = 0,
    method = "loess") + labs(title = "Data1")

plot_data2 <- ggplot(aes(x, y), data = data2) + geom_point() + geom_smooth(method = "lm",
    formula = y ~ x) + stat_smooth(colour = "firebrick", alpha = 0, method = "loess") +
    labs(title = "Data2")

plot_data3 <- ggplot(aes(x, y), data = data3) + geom_point() + geom_smooth(method = "lm",
    formula = y ~ x) + stat_smooth(colour = "firebrick", alpha = 0, method = "loess") +
    labs(title = "Data3")

plot_data4 <- ggplot(aes(x, y), data = data4) + geom_point() + geom_smooth(method = "lm",
    formula = y ~ x) + labs(title = "Data4")

grid.arrange(plot_data1, plot_data2, plot_data3, plot_data4, nrow = 2)
```



We can see that it is not appropriate to use linear regression for each of the graphs. The blue line on the graphs is the line from the `lm` function in r for generating the linear regression. The red line is generated by the `stat_smooth` function and method `loess` from `ggplot2` which gives us a locally weighted scatterplot smoothing [1].

Looking at each graph we can see where linear regression is appropriate and where it is not. `Data1` shows where linear regression is most appropriate and the scatter plot has a clear positive correlation. `Data2` is clearly a nonlinear regression as we can see the relationship has a diminishing slope over time exhibiting the curve. `Data3` does exhibit a scatterplot where linear regression is appropriate but we have one significant outlier that is impacting the linear regression significantly, the outlier may need to be removed or evaluted further. Lastly, `Data4` shows that linear regression would be totally inappropriate in this circumstance and infact I was unable to even fit a LOESS regression to this data set and graphically display it. `Data4` appears to be the worst data set to try an attempt fitting a linear regression.

It is important to visualize the data set when analyzing because we can spot issues quicker when they are visually displayed. As I worked through `Part II` of the exam I was completely unaware how different each data set was until I graphed them. All the evaluations I completed beforehand did not show a significant difference between the data sets until they were graphed together and now we can see very large and real differences between the data sets. The importance of this cannot be stressed enough, had I completed a lengthly analysis and then attempted to graph the data at the last part of the analysis many of my assumptions and testing would not be appropriate because the data sets are so different and are explaning very different types of observations.

References

[1] (n.d.). Retrieved December 14, 2015, from https://en.wikipedia.org/wiki/Local_regression